

PairCloneTree: Reconstruction of Tumor Subclone Phylogeny Based on Mutation Pairs using Next Generation Sequencing Data

Tianjian Zhou^{*1}, Subhajit Sengupta^{*2}, Peter Müller^{†3}, and Yuan Ji^{‡2, 4}

¹*Department of Statistics and Data Sciences, The University of Texas at Austin*

²*Program for Computational Genomics and Medicine, NorthShore University HealthSystem*

³*Department of Mathematics, The University of Texas at Austin*

⁴*Department of Public Health Sciences, The University of Chicago*

March 14, 2017

Abstract

We present a latent feature allocation model to reconstruct tumor subclones subject to phylogenetic evolution that mimics tumor evolution. Similar to most current methods, we consider data from next-generation sequencing. Unlike most methods that use information in short reads mapped to single nucleotide variants (SNVs), we consider subclone reconstruction using pairs of two proximal SNVs that can be mapped by the same short reads. As part of the Bayesian inference model, we construct a phylogenetic tree prior. The use of the tree structure in the prior greatly strengthens inference. Only subclones that can be approximated by a phylogenetic tree are assigned non-negligible probability. The proposed Bayesian framework implies posterior distributions on the number of subclones, their genotypes, cellular proportions, and the phylogenetic tree spanned by the inferred subclones. The proposed method is validated against different sets of simulated and real-world data using single and multiple tumor samples. An open source software package is available at <http://www.compgenome.org/pairclonetree>.

Keywords: Latent feature model; Mutation pair; NGS data; Phylogenetic tree; Subclone; Tumor heterogeneity

^{*}Have equal contribution

[†]Email: pmueller@math.utexas.edu

[‡]Email: yji@health.bsd.uchicago.edu

1 Introduction

Tumor cells emerge from a Darwinian-like selection among multiple competing subpopulations of cells [Nowell (1976); Bonavia et al. (2011); Marusyk et al. (2012)]. During tumorigenesis, through sequential clonal expansion and selection cells acquire distinct mutations. This process leads to genetically divergent subpopulations of cells, also known as subclones [Navin et al. (2010); Gerlinger et al. (2012); Nik-Zainal et al. (2012); Bignell et al. (2010); Bozic et al. (2010); Raphael et al. (2014)]. Reconstructing the subclones and their evolutionary relationship could help investigators to identify driver mutations that emerge early in the development or during the progression period. Such results provide insight about targeted therapies [Aparicio and Caldas (2013); Papaemmanuil et al. (2011); Varela et al. (2011); Stephens et al. (2012)].

A recent surge of genetic sequencing data makes it possible to investigate tumor subclonal architecture in detail [Oesper et al. (2013); Strino et al. (2013); Fischer et al. (2014); Miller et al. (2014); Roth et al. (2014); Jiao et al. (2014); Deshwar et al. (2015); Zare et al. (2014); Sengupta et al. (2015); Marass et al. (2017); Zhou et al. (2017)]. We will discuss details of some [Marass et al. (2017); Jiao et al. (2014); Deshwar et al. (2015)] later in Section 4, after we have introduced the required notation. Latest developments of next generation sequencing (NGS) technology enabled researchers to develop a variety of techniques that are broadly known as subclonal reconstruction. One of the aims is to deconvolute observed genomic data from a tumor into constituent signals corresponding to various subclones and to reconstruct their relationship in a phylogeny. In most methods the reconstruction is based on short reads that are mapped to single nucleotide variants (SNVs) (few methods also consider somatic copy number aberrations, SCNA). SNV-based subclone calling methods utilize variant allele fractions (VAFs), that is, the fractions of alleles (or short reads) at each locus that carry mutations. Since humans are diploid, the VAFs of short reads for a homogeneous cell population should be 0, 0.5 or 1.0 for any locus in copy number neutral (copy number = 2) regions and after adjusting for tumor purity. VAFs different from 0, 0.5 or 1.0 are therefore evidence for heterogeneity. Based on this idea, existing SNV-based subclone calling methods either cluster mutations [Miller et al. (2014); Roth et al. (2014); Jiao et al. (2014); Deshwar et al. (2015)], or use latent feature allocation methods to infer the subclone genotypes and their proportions [Zare et al. (2014); Sengupta et al. (2015); Marass et al. (2017)]. All are based on observed VAFs.

Main idea

We assume that the available data are from T ($T \geq 1$) samples from a single patient and the main inference goal is intra-tumor heterogeneity. We present a novel approach to reconstruct tumor subclones and their corresponding phylogenetic tree based on mutation pairs. Here a mutation pair refers to a pair of proximal SNVs on the genomes that can be simultaneously mapped by the same paired-end short reads, with one SNV on each end. In other words, mutation pairs can be phased by short reads. See Fig. 1 for an illustration. Short reads mapped to only one of the SNV loci are treated as partially missing paired-end reads and are not excluded from our approach. Specifically, marginal SNV reads can be included in our analysis. See Section 2.2 for more details. The idea of working with phased mutation pairs was introduced in Zhou et al. (2017). We build on this work and develop a novel and entirely different inference approach by explicitly modeling the underlying phylogenetic relationship. That is, we model tumor heterogeneity based on a representation of a phylogenetic tree of tumor cell subpopulations. A prior probability model on such phylogenetic trees induces a dependent prior on the mutation profiles of latent tumor cell subpopulations. In particular, the phylogenetic tree of tumor cell subpopulations is included as a random quantity in the Bayesian model. Currently, we only consider mutation pairs in copy neutral region i.e. copy number two. The proposed inference aims to reconstruct (i) subclones defined by the haplotypes across all the mutation pairs, (ii) cellular proportion of each subclone, and (iii) a phylogenetic tree spanned by the subclones.

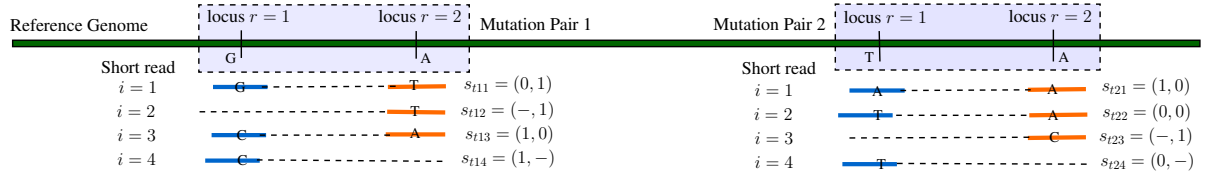


Figure 1: Short reads data from mutation pairs using NGS. Here s_{tki} denotes the i -th read for the k -th mutation pair in sample t . Each s_{tki} is a 2-dimensional vector which corresponds to the two proximal SNVs in the mutation pair, and each component of the vector takes values 0, 1 or $-$ representing wild type, variant or missing genotype, respectively.

Consider an NGS data set with K mutation pairs shared across all T ($T \geq 1$) samples. We assume that the samples are composed of C homogeneous subclones. The number of subclones C is unknown and becomes one of the model parameters. We use a $K \times C$ matrix \mathbf{Z} to represent the subclones, in which each column of \mathbf{Z} represents a subclone and each row represents a mutation pair. That is, the (kc) element z_{kc} of the matrix

corresponds to subclone c and mutation pair k . Each \mathbf{z}_{kc} is itself again a matrix. It is a 2×2 matrix that represents the genotypes of the two alleles of the mutation pair. See Fig. 2(b). An important step in the model construction is that the columns (subclones) of \mathbf{Z} form a phylogenetic tree \mathcal{T} . The tree encodes the parent-child relationship across the subclones. A detailed construction of the tree and a prior probability model of \mathcal{T} and \mathbf{Z} are introduced later. Lastly, we denote $\mathbf{w}_t = (w_{t1}, \dots, w_{tC})$ the cellular proportions of the subclones in sample t where $0 < w_{tc} < 1$ for all c and $\sum_{c=0}^C w_{tc} = 1$.

Using NGS data we infer \mathcal{T} , C , \mathbf{Z} and \mathbf{w} based on a simple idea that variant reads can only arise from subclones with variant alleles consistent with an underlying phylogeny. We develop a treed *latent feature allocation model* (LFAM) to implement this reconstruction. Mutation pairs are the objects of the LFAM, and subclones are the latent features chosen by the mutation pairs (in contrast to the phylogenetic Indian Buffet Process [Miller et al. (2008)] which builds a tree structure on objects, rather than features). Note that subclone reconstruction based on LFAM allows overlapping mutations across subclones and therefore does not require the infinite sites assumption [Nik-Zainal et al. (2012)]. This is different from many existing cluster-based models in the literature. While LFAM attempts to directly infer genotypes of all subclonal genomes, cluster-based models first infer SNV clusters based on VAFs and then reconstruct subclonal genotypes based on the clusters.

Advantage of using mutation pairs. Mutation pairs contain phasing information that improves the accuracy of subclone reconstruction. If two nucleotides reside on the same short read, we know that they must appear in the same DNA strand in a subclone. For example, consider a scenario with one mutation pair and two subclones. Suppose the reference genome allele is (A, G) for that mutation pair, with the notion that A and G are phased by the same DNA strand. Suppose the two subclones have diploid genomes at the two loci and the genotypes for both DNA strands are ((C, G), (A, T)) for subclone $c = 1$, and ((C, T), (A, G)) for $c = 2$. Since in NGS short reads are generated from a single DNA strand, short reads could be any of the four haplotypes (C, G), (A, T), (C, T) or (A, G) for this mutation pair. If indeed relative large counts of short reads with each haplotype are observed, one can reliably infer that there are heterogeneous cell subpopulations in the tumor sample. In contrast, if we ignore the phasing information and only consider the (marginal) VAFs for each SNV, then the observed VAFs for both SNVs are 0.5, which could be heterogeneous mutations from a single cell population. In this paper, we leverage the power of using mutation pairs over single SNVs to incorporate partial phasing information in our model. We assume that mutation pairs and their mapped short reads counts have been obtained using a bioinformatics pipeline, such as LocHap

[Sengupta et al. (2016)]. Our aim is to use short reads mapping data on mutation pairs to reconstruct tumor subclones and their phylogeny.

Difference from traditional phylogenetic tree. Phylogenetic trees are usually used to approximate perfect phylogeny for a fixed number of haplotypes [Gusfield (1991); Bafna et al. (2003); Pe’er et al. (2004)]. Most methods lack assessment of tree uncertainties and report a single tree estimate. Also, methods based on SNVs put the observed mutation profile of SNV at the leaf nodes. This is natural if the splits in the tree create subpopulations that acquire or do not acquire a new mutation (or set of mutations). In contrast, we define a tree with all descendant nodes differing from the parent node by some mutations. That is, all node, including interior nodes, correspond to tumor cell subpopulations. See details below. For clarification we note that the prior structure in our model is different from the phylogenetic Indian Buffet Process (pIBP) [Miller et al. (2008)], which models phylogeny of the objects rather than the features.

The rest of the paper is organized as follows: Section 2 and Section 3 describe the latent feature allocation model and posterior inference, respectively. Section 4 presents two simulation studies. Section 5 reports analysis results for an actual experiment. We conclude with a discussion in Section 6.

2 Statistical Model

2.1 Representation of Subclones

Fig. 2 presents a stylized example of temporal evolution of a tumor, starting from time T_0 and evolving until time T_4 with the normal clone (subclone $c = 1$) and three tumor subclones ($c = 2, 3, 4$). Each tumor subclone is marked by two mutation pairs with distinct somatic mutation profiles. In Fig. 2 the true phylogenetic tree is plotted connecting the stylized subclones. The true population frequencies of the subclones are marked in parentheses. In panel (b) subclone genomes, their population frequencies and the phylogenetic relationship are represented by \mathbf{Z} , \mathbf{w} , and \mathcal{T} . The entries of \mathcal{T} report for each subclone the index of the parent subclone (with $\mathcal{T}_1 = 0$ for the root clone $c = 1$).

Suppose K mutation pairs with C subclones are present. The subclone phylogeny can be visualized with a rooted tree with C nodes. We use a C -dimensional *parent* vector \mathcal{T} to encode the parent-child relationship of a tree, where $\mathcal{T}_c = \mathcal{T}[c] = j$ means that subclone j is the parent of subclone c . The parent vector uniquely defines the topology of a rooted tree. We assume that the tumor evolutionary process always starts from the

normal clone, indexed by $c = 1$. The normal clone does not have a parent, and we denote it by $\mathcal{T}_1 = 0$. For example, the parent vector representation of the subclone phylogeny in Fig. 2 is $\mathcal{T} = (0, 1, 1, 2)$.

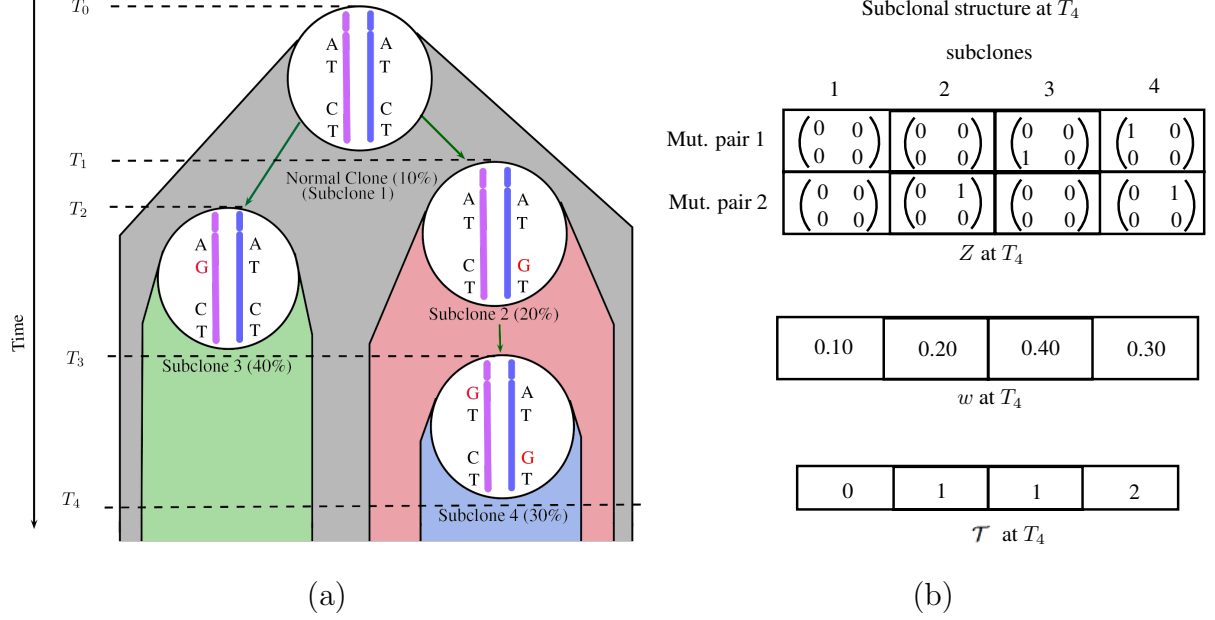


Figure 2: Schematic of subclonal evolution and subclone structure. Panel (a) shows the evolution of subclones over time. Panel (b) shows the subclonal structure at T_4 with genotypes \mathbf{Z} , cellular proportions \mathbf{w} and parent vector \mathcal{T} . For each mutation pair k and subclone c , the entry \mathbf{z}_{kc} of \mathbf{Z} is a 2×2 matrix corresponding to the arrangement in the figure in panel (a), that is, with alleles in the two columns, and SNVs in the rows.

We use the $K \times C$ matrix \mathbf{Z} to represent the subclone genotypes. Each column of \mathbf{Z} defines a subclone, and each row of \mathbf{Z} corresponds to a mutation pair. The entry \mathbf{z}_{kc} records the genotypes for mutation pair k in subclone c . Since each subclone has two alleles $j = 1, 2$, and each mutation pair has two loci $r = 1, 2$, the entry \mathbf{z}_{kc} is itself a 2×2 matrix, i.e. $\mathbf{z}_{kc} = (\mathbf{z}_{kcj}, j = 1, 2)$ and $\mathbf{z}_{kcj} = (z_{kcjr}, r = 1, 2)$,

$$\mathbf{z}_{kc} = (\mathbf{z}_{kc1}, \mathbf{z}_{kc2}) = \left[\begin{pmatrix} z_{kc11} \\ z_{kc12} \end{pmatrix} \begin{pmatrix} z_{kc21} \\ z_{kc22} \end{pmatrix} \right]$$

where $\begin{pmatrix} z_{kc11} \\ z_{kc12} \end{pmatrix}$ and $\begin{pmatrix} z_{kc21} \\ z_{kc22} \end{pmatrix}$ represent mutation pairs of allele 1 and allele 2, respectively. Theoretically, each z_{kcjr} can be any one of the four nucleotide sequences, A, C, G, T. However, at a single locus, the probability of having more than two sequences is negligible since it would require the same locus to be mutated twice throughout the life span of the tumor, which is extremely unlikely. Therefore, we assume z_{kcjr} can only take

two possible values, with $z_{kcjr} = 1$ (or 0) indicating that the corresponding locus has a mutation (or does not have a mutation) compared to the reference genome, respectively. For example, in Fig. 2, we have $K = 2$ mutation pairs and $C = 4$ subclones. For mutation pair $k = 2$ in subclone $c = 4$, the allele $j = 1$ harbors no mutation, while the allele $j = 2$ has a mutation at the first locus $r = 1$, which translates to $\mathbf{z}_{24} = (00, 10)$ (writing 00 as a shorthand for $(0, 0)^T$, etc.). Altogether, \mathbf{z}_{kc} can take $2^4 = 16$ possible values $\mathbf{z}_{kc} \in \{(00, 00), (00, 01), \dots, (11, 11)\}$. Since we do not have phasing information across mutation pairs, the \mathbf{z}_{kc} values having mirrored columns lead to exactly the same data likelihood and thus are indistinguishable. Therefore, we reduce the number of possible values of \mathbf{z}_{kc} to $Q = 10$. We list them below for further reference:

$\mathbf{z}^{(1)} = (00, 00)$, $\mathbf{z}^{(2)} = (00, 01)$, $\mathbf{z}^{(3)} = (00, 10)$, $\mathbf{z}^{(4)} = (00, 11)$, $\mathbf{z}^{(5)} = (01, 01)$, $\mathbf{z}^{(6)} = (01, 10)$, $\mathbf{z}^{(7)} = (10, 10)$, $\mathbf{z}^{(8)} = (01, 11)$, $\mathbf{z}^{(9)} = (10, 11)$ and $\mathbf{z}^{(10)} = (11, 11)$.

We assume that the normal subclone has no mutation, $\mathbf{z}_{k1} = \mathbf{z}^{(1)}$ for all k , indicating all mutations are somatic. In addition to these C true subclones, we introduce a background subclone, indexed as $c = 0$ and without biological meaning, to account for experimental noise and tiny subclones that are not detectable given the sequencing depth. We assume that the background subclone is a random mixture of all possible genotypes. See more discussion in Section 2.2.

Finally, we introduce notation for mixing proportions. Suppose T tissue samples are dissected from the same patient. We assume that the samples are admixtures of C subclones, each sample with a different set of mixing proportions (population frequencies). We use a $T \times (C + 1)$ matrix \mathbf{w} to record the proportions, where w_{tc} represents the population frequencies of subclone c in sample t , $0 < w_{tc} < 1$ and $\sum_{c=0}^C w_{tc} = 1$. The proportions w_{t1} denotes the proportion of normal cells contamination in sample t .

2.2 Sampling Model

Let \mathbf{N} be a $T \times K$ matrix with N_{tk} representing read depth for mutation pair k in sample t . It records the number of times any locus of the mutation pair is covered by sequencing reads (see Fig. 1). Let $\mathbf{s}_{tki} = (s_{tkir}, r = 1, 2)$ be a specific short read where $r = 1, 2$ index the two loci in a mutation pair, $i = 1, 2, \dots, N_{tk}$. We use $s_{tkir} = 1$ (or 0) to denote a variant (reference) sequence at the read, compared to the reference genome. An important feature of the data is that read i may not overlap with locus r . We use $s_{tkir} = -$ to represent the missing sequence on the read. Reads that do not overlap with either of the two loci are not included in the model as they do not contribute any information about the mutation pair. In summary, \mathbf{s}_{tki} can take $G = 8$ possible values,

$$\mathbf{s}_{tki} \in \{\mathbf{s}^{(1)}, \dots, \mathbf{s}^{(8)}\} = \{00, 01, 10, 11, -0, -1, 0-, 1-\}.$$

Among all N_{tk} reads, let $n_{tkg} = \sum_i I(\mathbf{s}_{tki} = \mathbf{s}^{(g)})$ be the number of short reads having genotype $\mathbf{s}^{(g)}$. As illustrated in Fig. 1 out of total 4 reads ($N_{t1} = 4$), we have $n_{t12} = 1, n_{t13} = 1, n_{t16} = 1$ and $n_{t18} = 1$.

We assume a multinomial sampling model for the observed read counts

$$(n_{tk1}, \dots, n_{tk8}) \mid N_{tk} \sim \text{Mn}(N_{tk}; p_{tk1}, \dots, p_{tk8}),$$

where p_{tkg} is the probability of observing a short read \mathbf{s}_{tki} with genotype $\mathbf{s}^{(g)}$. Later we link p_{tkg} with the underlying subclone structures.

If desired, it is straightforward to incorporate data for marginal SNV reads in the model. These reads can be treated as, without loss of generality, right missing reads, i.e. $s_{tki2} = -$. In this case, $n_{tk1} = \dots = n_{tk6} = 0$, and the multinomial sampling model reduces to a binomial model. The addition of marginal SNV counts does not typically improve inference. See more details in Zhou et al. (2017).

Construction of p_{tkg} . For a short read \mathbf{s}_{tki} , depending on whether it covers both loci or only one locus, we consider three cases: (i) a read covers both loci, taking values $\mathbf{s}_{tki} \in \{\mathbf{s}^{(1)}, \dots, \mathbf{s}^{(4)}\}$ (complete read); (ii) a read covers the second locus, taking values $\mathbf{s}_{tki} \in \{\mathbf{s}^{(5)}, \mathbf{s}^{(6)}\}$ (left missing read); and (iii) a read covers the first locus, taking values $\mathbf{s}_{tki} \in \{\mathbf{s}^{(7)}, \mathbf{s}^{(8)}\}$ (right missing read). Let $v_{tk1}, v_{tk2}, v_{tk3}$ denote the probabilities of observing a short read satisfying cases (i), (ii) and (iii), respectively. Conditional on cases (i), (ii) or (iii), let \tilde{p}_{tkg} be the conditional probability of observing $\mathbf{s}_{tki} = \mathbf{s}^{(g)}$. We have $p_{tkg} = v_{tk1} \tilde{p}_{tkg}, g = 1, \dots, 4$, $p_{tkg} = v_{tk2} \tilde{p}_{tkg}, g = 5, 6$, and $p_{tkg} = v_{tk3} \tilde{p}_{tkg}, g = 7, 8$. We assume non-informative missingness and do not make inference on v 's, so they remain constants in the likelihood.

We express \tilde{p}_{tkg} in terms of \mathbf{Z} and \mathbf{w} based on the following generative model. Consider a sample t . To generate a short read, we first select a subclone c with probability w_{tc} . Next we select with probability 0.5 one of the two alleles $j = 1, 2$. Finally, we record the read $\mathbf{s}^{(g)}$, $g = 1, 2, 3$ or 4 , corresponding to the chosen allele $\mathbf{z}_{kcj} = (z_{kcj1}, z_{kcj2})$. In the case of left (or right) missing locus we observe $\mathbf{s}^{(g)}$, $g = 5$ or 6 (or $g = 7$ or 8), corresponding to the observed locus of the chosen allele. Reflecting these three generative steps, we denote the probability of observing a short read $\mathbf{s}^{(g)}$ from subclone c that bears sequence \mathbf{z}_{kcj} by

$$A(\mathbf{s}^{(g)}, \mathbf{z}_{kc}) = \sum_{j=1}^2 0.5 \times I(s_1^{(g)} = z_{kcj1}) I(s_2^{(g)} = z_{kcj2}), \quad (1)$$

with the understanding that $I(- = z_{kcjr}) \equiv 1$ for missing reads. Implicit in (1) is the restriction $A(\mathbf{s}^{(g)}, \mathbf{z}_{kc}) \in \{0, 0.5, 1\}$, depending on the arguments.

Finally, using the conditional probabilities $A(\cdot)$ we obtain the marginal probability of observing a short read $\mathbf{s}^{(g)}$ from the tumor sample t with C subclones with cellular proportions $\{w_{tc}\}$ as

$$\tilde{p}_{tkg} = \sum_{c=1}^C w_{tc} A(\mathbf{s}^{(g)}, \mathbf{z}_{kc}) + w_{t0} \rho_g. \quad (2)$$

The first term in Eq. 2 states that the probability of observing a short read with genotype $\mathbf{s}^{(g)}$ is a weighted sum of the A 's across all the subclones. Here $w_{t0}\rho_g$ stands for the probability of observing $\mathbf{s}^{(g)}$ due to random noise. It can be thought of as a background subclone with weight w_{t0} , which is a random mixture of four genotypes 00, 01, 10 and 11 with proportions ρ_g . We assume the random noise does not differ across different mutation pairs, thus ρ_g does not have an index k . Note that $\rho_1 + \dots + \rho_4 = \rho_5 + \rho_6 = \rho_7 + \rho_8 = 1$. Again, the background subclone (denoted by $c = 0$) has no biological meaning and is only used to account for noise and artifacts in the NGS data (sequencing errors, mapping errors, etc.).

2.3 Prior Model

We construct a hierarchical prior model, starting with $p(C)$, then a prior on the tree for a given number of nodes, $p(\mathcal{T} \mid C)$, and finally a prior on the subclonal genotypes given the phylogenetic tree \mathcal{T} .

Prior for C and \mathcal{T} . We assume a geometric prior for the number of subclones, $p(C) = (1 - \alpha)^{C-1}\alpha$, $C \in \{1, 2, 3, \dots\}$. Conditional on C , the prior on the tree, $p(\mathcal{T} \mid C)$ is as in [Chipman et al. \(1998\)](#). For a tree with C nodes, we let

$$p(\mathcal{T} \mid C) \propto \prod_{c=1}^C (1 + \eta_c)^{-\beta},$$

where η_c is the depth of node c , or the number of generations between node c and the normal subclone 1. The prior penalizes deeper trees and thus favors parsimonious representation of subclonal structure.

Prior for \mathbf{Z} . The subclone genotype matrix \mathbf{Z} can be thought of as a feature allocation for categorical matrices. The mutation pairs are the objects, and the subclones are the latent features chosen by the objects. Each feature has 10 categories corresponding to the $Q = 10$ different genotypes. Conditioning on \mathcal{T} the subclone genotype matrix needs to introduce dependence across features to reflect the assumed phylogeny. We construct a prior for \mathbf{Z} based on the following generative model. We start from a normal subclone denoted by $\mathbf{z}_{.1} = \mathbf{0}$. Now consider a subclone $c > 1$ and defined by $\mathbf{z}_{.c}$. The subclone

preserves all mutations from its parent $\mathbf{z}_{\mathcal{T}_c}$, but also gains a Poisson number of new mutations. We assume the new mutations randomly happen at the unmutated loci of the parent subclone. A formal description of prior of \mathbf{Z} follows.

For a subclone c , let $\ell_{kc} = \sum_{j,r} z_{kcjr}$ denote the number of mutations in mutation pair k , and let $\mathcal{L}_c = \{k : \ell_{kc} < 4\}$ denote the mutation pairs in subclone c that have less than four mutations. Let $m_{kc} = \ell_{kc} - \ell_{k\mathcal{T}_c}$ denote the number of new mutations that mutation pair k gains compared to its parent, and let $m_{\cdot c} = \sum_k m_{kc}$. We assume (i) The child subclone should acquire at least one additional mutation compared with its parent (otherwise subclone c would be identical to its parent \mathcal{T}_c). (ii) If the parent has already acquired all four mutations for a given k , then the child can not gain any more new mutation. That is, if $\ell_{k\mathcal{T}_c} = 4$, then $m_{kc} = 0$. (iii) Each mutation pair can gain at most one additional mutation in each generation, $m_{kc} \in \{0, 1\}$. Based on these assumptions, given a parent subclone $\mathbf{z}_{\mathcal{T}_c}$, we construct a child subclone $\mathbf{z}_{\cdot c}$ as follows. Let $\mathcal{M}_c = \{k : m_{kc} = 1\}$ be the set of mutation pairs in subclone c where new mutations are gained. Let $\text{Choose}(\mathcal{L}, m)$ denote a uniformly chosen subset of \mathcal{L} of size m , and let $X \sim \text{Trunc-Pois}(\lambda; [a, b])$ represent a Poisson distribution with mean λ , truncated to $a \leq X \leq b$. We assume

$$\begin{aligned} m_{\cdot c} \mid \mathbf{z}_{\mathcal{T}_c}, \mathcal{T}, C &\sim \text{Trunc-Pois}(\lambda; [1, |\mathcal{L}_{\mathcal{T}_c}|]), \\ \mathcal{M}_c \mid m_{\cdot c}, \mathbf{z}_{\mathcal{T}_c}, \mathcal{T}, C &\sim \text{Choose}(\mathcal{L}_{\mathcal{T}_c}, m_{\cdot c}). \end{aligned} \quad (3)$$

The lower bound and upper bound of the truncated Poisson reflect assumptions (i) and (ii) respectively. Also, Eq. 3 implicitly captures assumption (iii).

Next, for a mutation pair that gains one new mutation, we assume the new mutation randomly arises in any of the unmutated loci in the parent subclone. Let $\mathcal{Z}_{kc} = \{(j, r) : z_{kcjr} = 0\}$, and let $\text{Unif}(A)$ denote a uniform distribution over the set A . We first choose

$$(j^*, r^*) \mid \mathbf{z}_{\mathcal{T}_c}, \mathcal{T}, C \sim \text{Unif}(\mathcal{Z}_{k\mathcal{T}_c}),$$

and then set $z_{kcj^*r^*} = 1$. So we have

$$p(\mathbf{Z} \mid \mathcal{T}, C) \propto \prod_{c=2}^C \text{Trunc-Pois}(m_{\cdot c}; [1, |\mathcal{L}_{\mathcal{T}_c}|]) \cdot \frac{1}{\binom{|\mathcal{L}_{\mathcal{T}_c}|}{m_{\cdot c}}} \cdot \prod_{k \in \mathcal{M}_c} \frac{1}{|\mathcal{Z}_{k\mathcal{T}_c}|}.$$

Prior for \mathbf{w} and ρ . We design $p(\mathbf{w})$ in such a manner that we could put an informative prior for w_{t1} if a reliable estimate for tumor purity is available based on some prior bioinformatics pipeline (e.g. Van Loo et al. (2010); Carter et al. (2012)). Recall that $c = 1$ is the normal subclone, i.e., w_{t1} is the normal subclone proportion, and that

$\sum_{c=0, c \neq 1}^C w_{tc} + w_{t1} = 1$. We assume a Beta-Dirichlet prior on \mathbf{w} such that,

$$w_{t1} \sim \text{Be}(a_p, b_p); \quad \text{and} \quad \frac{w_{tc}}{(1 - w_{t1})} \sim \text{Dir}(d_0, d, \dots, d),$$

where $c = 0, 2, 3, \dots, C$. We set $d_0 \ll d$ as w_{t0} is only a correction term to account for background noise and model mis-specification term.

The model is completed with a prior for $\boldsymbol{\rho} = \{\rho_g\}$. We consider complete read, left missing read and right missing read separately, and assume

$$\rho_{g_1} \sim \text{Dir}(d_1, \dots, d_1); \quad \rho_{g_2} \sim \text{Dir}(2d_1, 2d_1); \quad \rho_{g_3} \sim \text{Dir}(2d_1, 2d_1),$$

where $g_1 = \{1, 2, 3, 4\}$, $g_2 = \{5, 6\}$ and $g_3 = \{7, 8\}$.

3 Posterior Inference

Let $\mathbf{x} = (\mathbf{Z}, \mathbf{w}, \boldsymbol{\rho})$ denote the unknown parameters except for the number of subclones C and the tree \mathcal{T} . Markov chain Monte Carlo (MCMC) simulation from the posterior $p(\mathbf{x} \mid \mathbf{n}, \mathcal{T}, C)$ is used to implement posterior inference. Gibbs sampling transition probabilities are used to update \mathbf{Z} , and Metropolis-Hastings transition probabilities are used to update \mathbf{w} and $\boldsymbol{\rho}$. For example, we update \mathbf{Z} by row with

$$p(\mathbf{z}_{k\cdot} \mid \mathbf{z}_{-k\cdot}, \dots) \propto \prod_{t=1}^T \prod_{g=1}^G \left[\sum_{c=1}^C w_{tc} A(\mathbf{h}_g, \mathbf{z}_{kc}) + w_{t0} \rho_g \right]^{n_{tkg}} \cdot p(\mathbf{z}_{k\cdot} \mid \mathbf{z}_{-k\cdot}, \mathcal{T}, C),$$

where $\mathbf{z}_{k\cdot}$ is a row of \mathbf{Z} satisfying the phylogeny \mathcal{T} .

Since the posterior distribution $p(\mathbf{x} \mid \mathbf{n}, \mathcal{T}, C)$ is expected to be highly multi-modal, we utilize parallel tempering [Geyer (1991)] to improve the mixing of the chain. Specifically, we use OpenMP parallel computing API [Dagum and Menon (1998)] in C++, to implement a scalable parallel tempering algorithm.

Updating C and \mathcal{T} . In general, posterior MCMC on tree structures can be very challenging to implement [Chipman et al. (1998); Denison et al. (1998)]. However, the problem here is manageable since plausible numbers for C constrain \mathcal{T} to moderately small trees. We assume that the number of nodes is *a priori* restricted to $C_{\min} \leq C \leq C_{\max}$. Conditional on the number of subclones C , the number of possible tree topologies is finite. Let \mathcal{T} denote the (discrete) sample space of (\mathcal{T}, C) . Updating the values of (\mathcal{T}, C) involves trans-dimensional MCMC. At each iteration, we propose new values for (\mathcal{T}, C) from a uniform proposal, $q(\tilde{\mathcal{T}}, \tilde{C} \mid \mathcal{T}, C) \sim \text{Unif}(\mathcal{T})$.

In order to search the space \mathcal{T} for the number of subclones and trees that best explain the observed data, we follow a similar approach as in Lee et al. (2015); Zhou et al. (2017)

(motivated by fractional Bayes' factor in O'Hagan (1995)) that splits the data into a training set and a test set. Recall that \mathbf{n} represents the read counts data. We split \mathbf{n} into a training set \mathbf{n}' with $n'_{tkg} = bn_{tkg}$, and a test set \mathbf{n}'' with $n''_{tkg} = (1 - b)n_{tkg}$. Let $p_b(\mathbf{x} \mid \mathcal{T}, C) = p(\mathbf{x} \mid \mathbf{n}', \mathcal{T}, C)$ be the posterior evaluated on the training set only. We use p_b in two instances. First, p_b is used as an informative prior instead of the original prior $p((\mathbf{x} \mid \mathcal{T}, C)$, and second, p_b is used as a proposal distribution for $\tilde{\mathbf{x}}$, $q(\tilde{\mathbf{x}} \mid \tilde{\mathcal{T}}, \tilde{C}) = p_b(\tilde{\mathbf{x}} \mid \tilde{\mathcal{T}}, \tilde{C})$. Finally, the acceptance probability of proposal $(\tilde{\mathcal{T}}, \tilde{C}, \tilde{\mathbf{x}})$ is evaluated on the test set. Importantly, in the acceptance probability the (intractable) normalization constant of p_b cancels out, making this approach computationally feasible.

$$p_{\text{acc}}(\mathcal{T}, C, \mathbf{x}, \tilde{\mathcal{T}}, \tilde{C}, \tilde{\mathbf{x}}) = 1 \wedge \frac{p(\mathbf{n}'' \mid \tilde{\mathbf{x}}, \tilde{\mathcal{T}}, \tilde{C})}{p(\mathbf{n}'' \mid \mathbf{x}, \mathcal{T}, C)} \cdot \frac{p(\tilde{\mathcal{T}}, \tilde{C}) \cancel{p_b(\tilde{\mathbf{x}} \mid \tilde{\mathcal{T}}, \tilde{C})}}{p(\mathcal{T}, C) \cancel{p_b(\mathbf{x} \mid \mathcal{T}, C)}} \cdot \frac{q(\mathcal{T}, C \mid \tilde{\mathcal{T}}, \tilde{C}) \cancel{q(\mathbf{x} \mid \mathcal{T}, C)}}{q(\tilde{\mathcal{T}}, \tilde{C} \mid \mathcal{T}, C) \cancel{q(\tilde{\mathbf{x}} \mid \tilde{\mathcal{T}}, \tilde{C})}}.$$

Here we use p_b as an informative proposal distribution for $\tilde{\mathbf{x}}$ to achieve a better mixing Markov chain Monte Carlo simulation with reasonable acceptance probabilities. Without the use of an informative proposal, the proposed new tree is almost always rejected because the multinomial likelihood with the large sample size is very peaked. Under the modified prior $p_b(\cdot)$, the resulting conditional posterior on \mathbf{x} remains entirely unchanged, $p_b(\mathbf{x} \mid \mathcal{T}, C, \mathbf{n}) = p(\mathbf{x} \mid \mathcal{T}, C, \mathbf{n})$ [Zhou et al. (2017)].

The described uniform tree proposal is in contrast to usual search algorithms for trees that generate proposals from neighboring trees. The advantage of this kind of proposal is to ensure a reasonable acceptance probability. But such algorithms have an important drawback that they quickly gravitate towards a local mode and then get stuck. A possible approach to addressing this problem is to repeatedly restart the algorithm from different starting trees. See Chipman et al. (1998) for more details. Our uniform tree proposal combined with the data splitting scheme is another way to mitigate this challenge, efficiently searching the tree space while keeping a reasonable acceptance probability.

Point estimates for parameters. All posterior inference is contained in the posterior distribution for \mathbf{x} , C and \mathcal{T} . For example, the marginal posterior distribution of C and \mathcal{T} gives updates posterior probabilities for all possible values of C and \mathcal{T} . It is still useful to report point estimates. We use the posterior modes $(\hat{C}, \hat{\mathcal{T}})$ as point estimates for (C, \mathcal{T}) , and conditional on \hat{C} and $\hat{\mathcal{T}}$, we use the maximum a posteriori (MAP) estimator as an estimation for the other parameters. The MAP is approximated as the MCMC sample with highest posterior probability. Let $\{\mathbf{x}^{(l)}, l = 1, \dots, L\}$ be a set of MCMC samples of

\mathbf{x} , and

$$\hat{l} = \arg \max_{l \in \{1, \dots, L\}} p(\mathbf{n} \mid \mathbf{x}^{(l)}, \hat{\mathcal{T}}, \hat{C}) p(\mathbf{x}^{(l)} \mid \hat{\mathcal{T}}, \hat{C}).$$

We report point estimates as $\hat{\mathbf{Z}} = \mathbf{Z}^{(\hat{l})}$, $\hat{\mathbf{w}} = \mathbf{w}^{(\hat{l})}$ and $\hat{\boldsymbol{\rho}} = \boldsymbol{\rho}^{(\hat{l})}$.

4 Simulation Studies

We present two simulation studies to assess the proposed approach. We simulate single sample and multi-sample data with different read depths to test the performance of our model in different scenarios. In both simulation studies, we generate hypothetical read count data for $K = 100$ mutation pairs, which is a typical number of mutation pairs in a tumor sample. However, if needed, a much larger number of SNVs could be included in the model, with the only limiting concern being computational efficiency, which remains a challenge for all current methods.

4.1 Simulation 1

In the first simulation study, we consider $T = 1$ sample, which is the case for most real-world tumor cases due to the challenge in obtaining multiple samples from a patient. However, this does not rule out meaningful inference. As we will show, with good read depth, the simulation truth can still be recovered. Note that the relevant sample size is not the number of tissue samples, but closer to the number of reads, which is large even for $T = 1$.

We consider $K = 100$ mutation pairs and assume a simulation truth with $C = 4$ latent subclones. Fig. 3(a) and (d) show the true underlying subclonal genotypes and phylogeny, respectively. We use a heatmap to show the subclone matrix \mathbf{Z} , where colors from light gray to red to black are used to represent genotypes $\mathbf{z}^{(1)}$ to $\mathbf{z}^{(10)}$. The subclone weights are simulated from $\text{Dir}(0.01, \sigma(15, 10, 8, 5))$, where $\sigma(15, 10, 8, 5)$ stands for a random permutation of the four numbers. For the single sample in this simulation we get $\mathbf{w} = (0.000, 0.135, 0.169, 0.470, 0.226)$. The noise factor $\boldsymbol{\rho}$ is generated from its prior with $d_1 = 1$. In order to mimic a typical rate of observing left (or right) missing reads, we set $v_{tk2} = v_{tk3} = 0.25$, for $k = 1, \dots, 50$, and $v_{tk2} = v_{tk3} = 0.3$, for $k = 51, \dots, 100$. For the read depth N_{tk} , we consider two scenarios. In the first scenario, we consider 500x depth and generate $N_{tk} \sim \text{Discrete-Unif}([400, 600])$; in the second scenario, we consider 2000x depth and generate $N_{tk} \sim \text{Discrete-Unif}([1900, 2100])$. While these read depth values are impossible from existing whole-genome sequencing technology, they are available from whole-exome or targeted sequencing experiments.

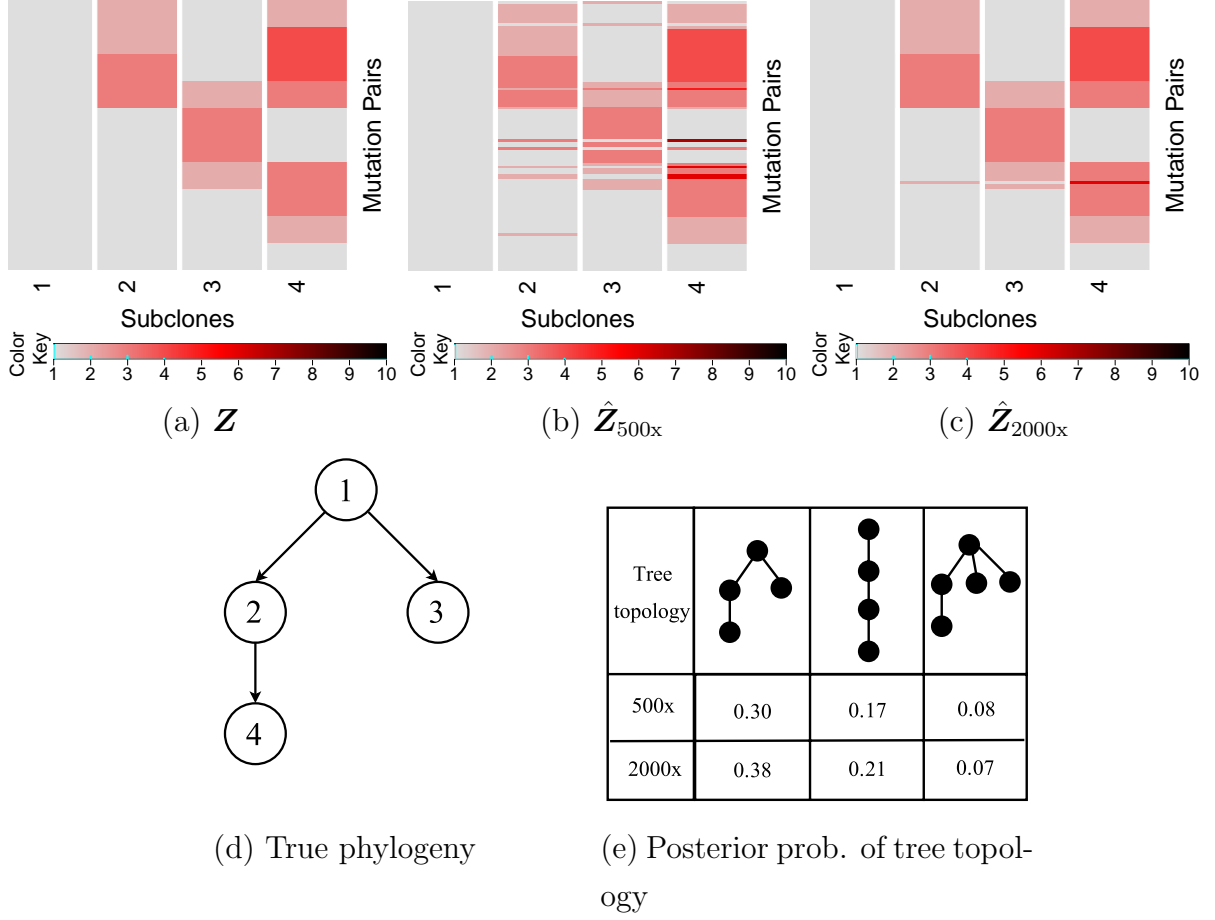


Figure 3: Simulation 1. Simulation truth \mathbf{Z} (a) and phylogeny (d), and posterior inference under PairCloneTree (b, c, e).

We fit the model with the following hyperparameters: $\alpha = 0.5$, $\beta = 0.5$, $d = 0.5$, $d_0 = 0.03$, $d_1 = 1$, where the values of α and β imply mild penalty for deep and bushy trees [Chipman et al. (1998)], and other hyperparameters are generic non-informative choices. We set $a_p = d$, $b_p = d_0 + (C - 1)d$ for given C as a non-informative prior choice and set $\lambda = 2K/C$ to express our prior belief that about half of the mutations occur uniformly at each generation. We use $C_{\min} = 2$ and $C_{\max} = 5$ as the range of C , since the vast majority of the methods in the literature show that even though a tumor sample could possess thousands to millions of SNVs, the number of inferred subclones usually is in the low single digit. Empirically, we choose the training set fraction as $b = 0.95$, as it performs well in all simulation studies. We run a total of 8000 MCMC simulations. Discarding the first 3000 draws as initial burn-in, we have a Monte Carlo sample with 5000 posterior draws.

Posterior inference with 500x read depth is summarized in Fig. 3(b, e). Fig. 3(e) shows

the top three tree topologies and corresponding posterior probabilities. The posterior mode recovers the true phylogeny. Fig. 3(b) shows the estimated genotypes with 500x read depth, conditional on the posterior modes (\hat{C}, \hat{T}) . Some mismatches are due to the single sample and limited read depth. The estimated subclone proportions are $\hat{\mathbf{w}} = (0.000, 0.073, 0.171, 0.517, 0.239)$, which agrees with the truth.

Posterior inference with 2000x read depth is summarized in Fig. 3(c, e). The posterior mode recovers the true phylogeny. Fig. 3(c) shows the estimated genotypes. The simulation shows how larger read depths improve posterior accuracy and improve the power of recovering the latent structure. In particular, this shows that even with a single sample, with reasonable read depth, the truth can still be recovered. The estimated subclone proportions are $\hat{\mathbf{w}} = (0.000, 0.127, 0.168, 0.477, 0.228)$.

4.2 Comparison with Cloe and PhyloWGS

There is no other subclone calling method based on paired-end read data that also infers phylogeny. We therefore compare with other similar model-based approaches. In particular, we use Cloe [Marass et al. (2017)] and PhyloWGS [Jiao et al. (2014); Deshwar et al. (2015)] for inference with the same simulated data. These two methods also use highly structured Bayesian nonparametric priors and MCMC simulations for posterior inference. Both methods take mutant read counts and read depths for SNVs as input. Therefore, we discard the phasing information in mutation pairs and only record marginal counts for SNVs as the input. The simulation truth in Cloe and PhyloWGS’s format is shown in Fig. 4(a). The orange color means a heterozygous mutation at the corresponding SNV locus.

Cloe infers clonal genotypes and phylogeny based on a similar feature allocation model. We run Cloe with the default hyperparameters for the same number of 8000 iterations with the first 3000 draws as initial burn-in. After that we carry out model selection for C with $2 \leq C \leq 5$. For the 500x read depth data, based on MAP estimate, Cloe reports 3 subclones with phylogeny $1 \rightarrow 2 \rightarrow 3$, and the subclone genotypes are shown in Fig. 4(b) with subclone proportions $\hat{\mathbf{w}}^{\text{Cloe}} = (0.569, 0.218, 0.213)$. For the 2000x read depth data, Cloe infers 2 subclones (genotypes not shown).

PhyloWGS, on the other hand, infers clusters of mutations and phylogeny. One can then make phylogenetic analysis to conjecture subclones and genotypes. Let $\tilde{\phi}_i$ denote the fraction of cells with a variant allele at locus i . The $\tilde{\phi}_i$ ’s are latent quantities related to the observed VAF for each SNV. PhyloWGS infers the phylogeny by clustering SNVs with matching $\tilde{\phi}_i$ ’s under a tree-structured prior for the unique values ϕ_j . In particular, they use the tree-structured stick breaking process (TSSB) [Adams et al. (2010)]. The

TSSB implicitly defines a prior on the formation of subclones, including the prior on C and the number of novel loci that arise in each subclone. In contrast, PairCloneTree explicitly defines these model features, allowing easier prior control on C and \mathcal{M}_c . We run PhyloWGS with the default hyperparameters and 2500 iterations with a burn-in of 1000 samples. We only consider loci with VAF > 0 as the other loci do not provide information for PhyloWGS clustering. We then report cluster sizes and phylogeny based on MAP estimate. For the 500x read depth data, PhyloWGS reports 3 subclones with phylogeny $0 \rightarrow 1(77, 0.429) \rightarrow 2(53, 0.218)$, where 0 refers to the normal subclone, and the numbers in the brackets refer to the cluster sizes and cellular prevalences. The conjectured subclone genotypes are shown in Fig. 4(c), with subclone proportions $\hat{\mathbf{w}}^{\text{PWGS}} = (0.571, 0.211, 0.218)$. For the 2000x read depth data, PhyloWGS reports 3 subclones with phylogeny $0 \rightarrow 1(80, 0.431) \rightarrow 2(50, 0.227)$ (genotypes not shown).

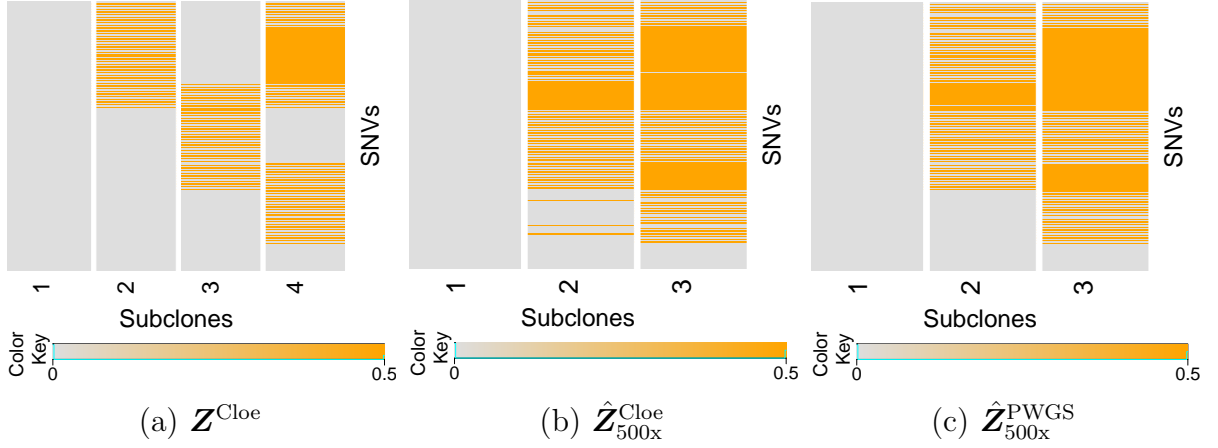


Figure 4: Simulation 1. Simulation truth \mathbf{Z}^{Cloe} (a), and posterior inference under Cloe (b) and PhyloWGS (c).

Inferences under Cloe and PhyloWGS do not entirely recover the truth. The reason is probably that the common mutations of subclones 2 and 4 (\mathcal{M}_2 with a cellular prevalence of $0.169 + 0.226$) have a similar cellular prevalence with the mutations of subclone 3 (\mathcal{M}_3 with a cellular prevalence of 0.470). Here we abuse the notation slightly and let \mathcal{M}_c denote the new mutations that subclone c gained. Therefore, Cloe infers that \mathcal{M}_2 and \mathcal{M}_3 belong to the same subclone ($\mathcal{M}_2^{\text{Cloe}} \approx \mathcal{M}_2 \cup \mathcal{M}_3$ and $\mathcal{M}_3^{\text{Cloe}} \approx \mathcal{M}_4$). Similarly, PhyloWGS clusters \mathcal{M}_2 and \mathcal{M}_3 together. Using more informative mutation pairs data, PairCloneTree is able to identify that \mathcal{M}_2 and \mathcal{M}_3 belong to different subclones. The comparisons support the argument in Section 1 that the inclusion of phasing information from the paired-end read data increases statistical power in recovering the underlying structure. Note that PairCloneTree is based on a different sampling model and has a

very different representation of \mathbf{Z} . Therefore, there is no obvious way to quantify the three model’s performance under the same scale.

4.3 Simulation 2

In the second simulation, we evaluate the performance of the proposed approach on multiple samples. We still consider $K = 100$ mutation pairs, but with a more complicated subclone structure, $C = 5$. We generate hypothetical data for $T = 8$ samples. The subclone proportions in each sample t are generated from $\mathbf{w}_t \sim \text{Dir}(0.01, \sigma(25, 15, 10, 8, 5))$. Fig. 5(a, b, c) show simulation truth \mathbf{Z} , \mathbf{w} and the phylogeny, respectively. We show \mathbf{w} in a heatmap with light gray to deep blue scale. A darker blue color indicates higher abundance of a subclone in a sample, while a lighter gray color indicates lower abundance. The proportions of the background subclone w_{t0} ’s are not shown as they only take tiny values, $w_{t0} < 10^{-3}$. The average sequence depth for the eight samples was about 500x.

The hyperparameters are set to be the same as in simulation 1. We run the same number of MCMC iterations.

The true phylogeny is recovered with 100% posterior probability (Fig. 5(c)). Fig. 5(d, e) show the estimated genotypes $\hat{\mathbf{Z}}$ and subclone proportions $\hat{\mathbf{w}}$. The truth is exactly recovered. The simulation shows that with more information from eight samples inference becomes quite reliable.

For comparison we again run Cloe and PhyloWGS on this data. Cloe correctly infers the number of subclones, and the estimated subclone genotypes match the truth, shown in Fig. 5(f). However, Cloe infers the phylogeny as $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5$. On the other hand, PhyloWGS infers the phylogeny as $0 \rightarrow 1 \begin{matrix} \rightarrow 2 \\ \rightarrow 3 \end{matrix}$ (details not shown). Both methods approximate but still miss some detail in the simulation truth.

5 Lung Dataset

We use whole-exome sequencing (WES) data generated from four ($T = 4$) surgically dissected tumor samples taken from a single patient diagnosed with lung adenocarcinoma. DNA is extracted from all four samples and exome library is sequenced on an Illumina HiSeq 2000 platform in paired-end fashion. Each of the read is 100 base-pair long and coverage is 200x-400x. We use BWA [Li and Durbin (2009)] and GATK’s UniformGenotyper [McKenna et al. (2010)] for mapping and variant calling, respectively. In order to find mutation pair location along with their genotypes with number of reads supporting them, we use a bioinformatics tool called LocHap [Sengupta et al. (2016)]. This tool

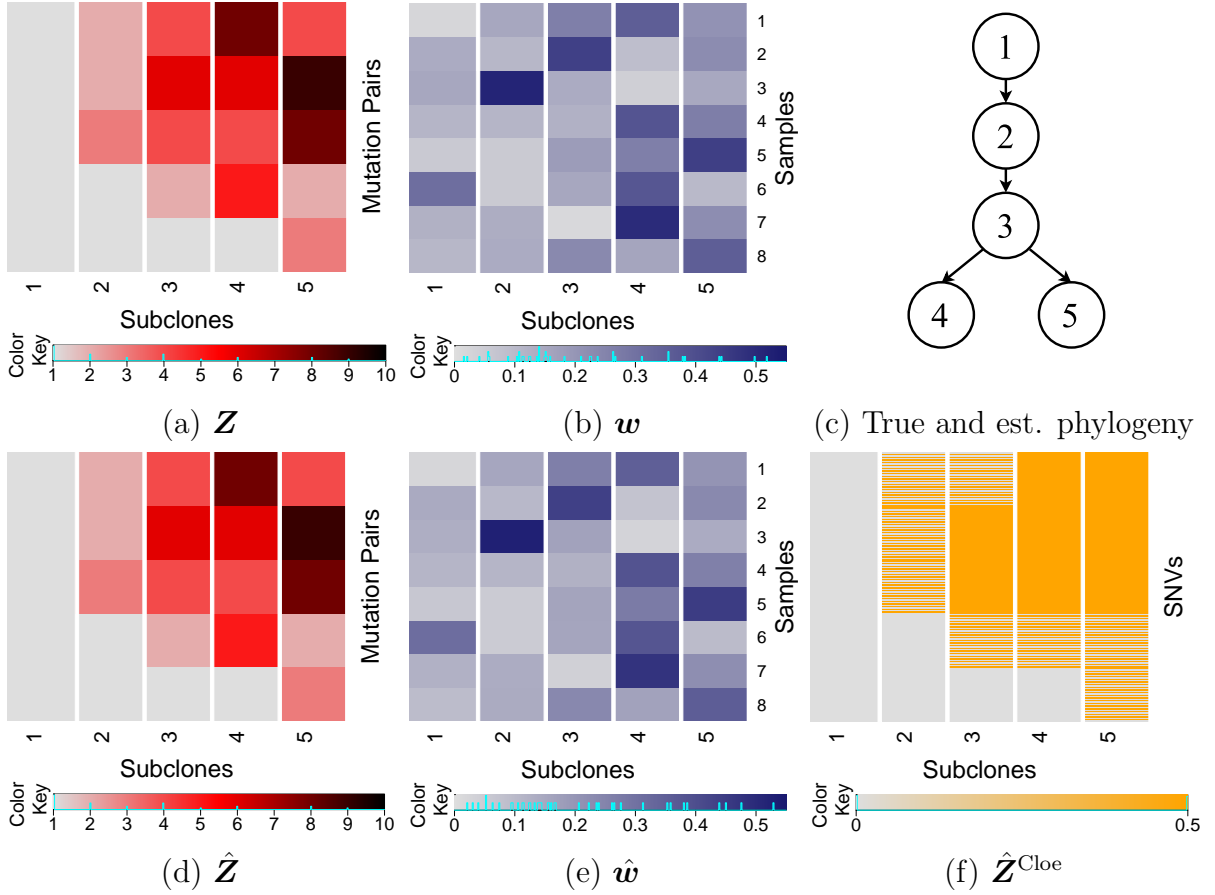


Figure 5: Simulation 2. Simulation truth \mathbf{Z} (a), \mathbf{w} (b) and phylogeny (c), and posterior inference under PairCloneTree (c, d, e) and Cloe (f).

searches for two or three SNVs that are scaffolded by the same reads. When the scaffolded SNVs, known as local haplotypes, exhibits more than two haplotypes, it is known as local haplotype variant (LHV). Using the individual BAM and VCF files **LocHap** finds a few hundreds LHVs on average in a WES sample. We select LHVs with two SNVs as we are interested in mutation pairs only. On those LHVs, we run the bioinformatics filters suggested by **LocHap** to keep the mutation pairs with high calling quality. We focus our analysis in copy number neutral regions. In the end, we get 69 mutation pairs for the sample and record the read count data from **LocHap**'s output.

We use the same hyperparameters and MCMC setting as in the simulations. Fig. 6 (d) shows some of the the posterior probabilities of the subclone phylogeny. The posterior mode is shown in Fig. 6 (c) with $C = 5$ subclones. Fig. 6 (a, b) show the estimated subclone genotypes $\hat{\mathbf{Z}}$ and cellular proportions $\hat{\mathbf{w}}$, respectively ($\hat{w}_{t0} < 4 \times 10^{-3}$ and are not shown). The rows for $\hat{\mathbf{Z}}$ are reordered for better display. The cellular proportions of the subclones show strong similarity across the 4 samples, indicating homogeneity of

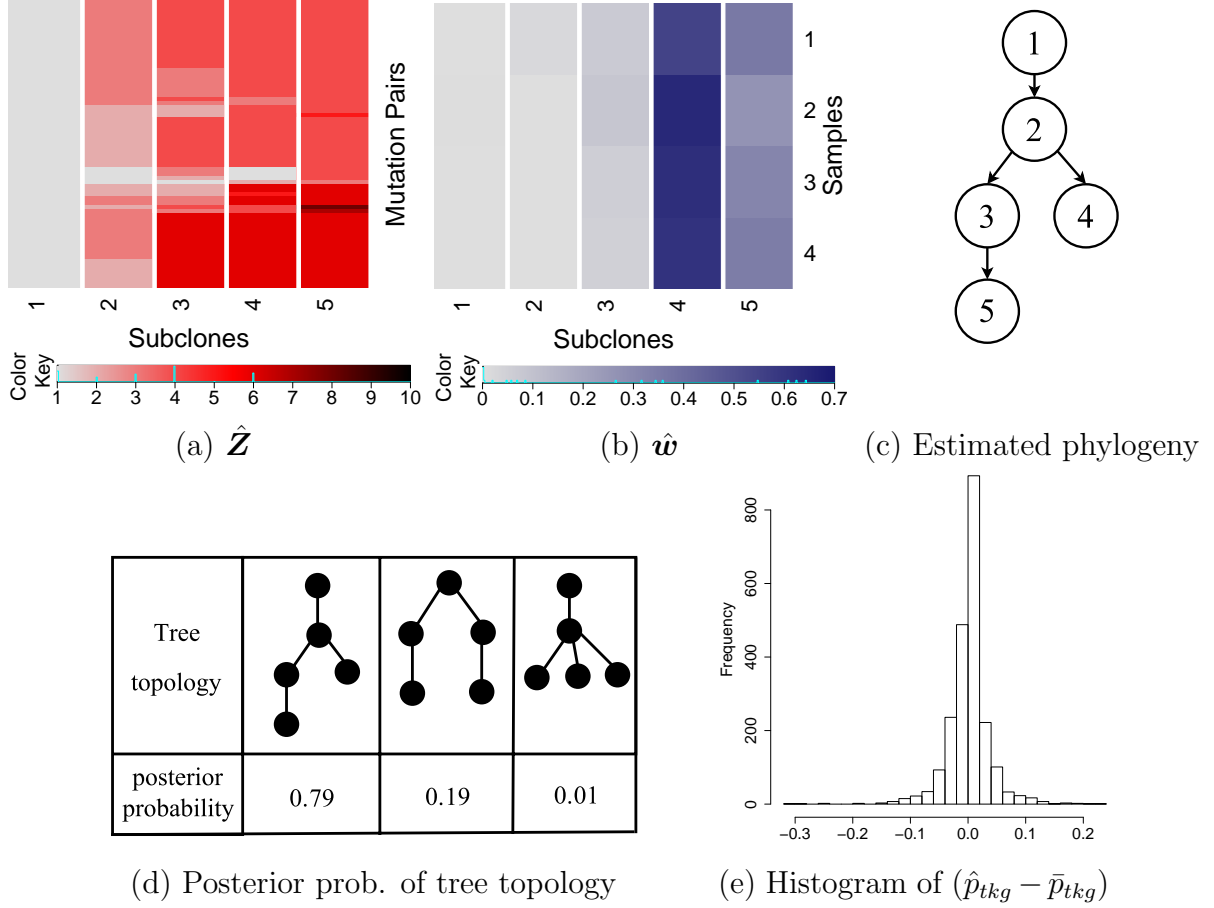


Figure 6: Posterior inference with PairCloneTree for lung cancer data set.

the samples. This is expected as the samples are dissected from proximal sites. Subclone 1, which is the normal subclone, takes a small proportion in all 4 samples, indicating high purity of the tumor samples. Subclones 2 and 3 are also included in only small proportions. They have almost vanished in the samples. However, as parents of subclones 4 and 5, respectively, they are important for the reconstruction of the subclone phylogeny. Subclones 4 and 5 are the two main subclones. They share a large proportion of common mutations, but each one has some private mutations, consistent with the estimated tree. Finally, Fig. 6 (e) shows a histogram of residuals, where we calculate empirical values $\bar{p}_{tkg} = n_{tkg}/N_{tk}$ and plot the difference $(\hat{p}_{tkg} - \bar{p}_{tkg})$. The residuals are centered at zero with little variation, indicating a good model fit.

For comparison, we run Cloe and PhyloWGS on the same data set with default hyperparameters. Cloe infers four subclones with phylogeny $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$. Fig. 7 (a, b) show the estimated genotypes $\hat{\mathbf{Z}}^{\text{Cloe}}$ and cellular proportions $\hat{\mathbf{w}}^{\text{Cloe}}$, respectively. PhyloWGS estimates 6 clusters (and a cluster 0 for normal subclone) of the SNVs with

phylogeny

$$\begin{array}{ccccccc} 0 & \rightarrow & 1 & \rightarrow & 2 & \rightarrow & 3 & \rightarrow & 4 \\ & & & & & & \rightarrow & 5 & \rightarrow & 6 \end{array} .$$

Fig. 7 (c) summarizes the cluster sizes and cellular prevalences. In light of the earlier simulation studies we believe that the inference under PairCloneTree is more reliable. Cloe and PhyloWGS outputs confirm that the four samples have similar proportions of all the subclones, indicating little inter-sample heterogeneity. Also, Cloe and PhyloWGS infer very small normal cell proportion, which corroborates PairCloneTree’s finding that the tumor samples have high purity.

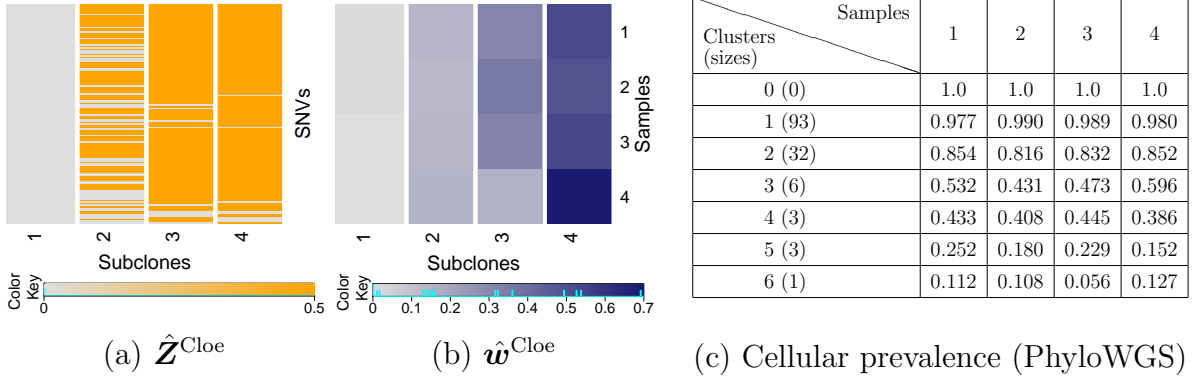


Figure 7: Posterior inference with Cloe (a, b) and PhyloWGS (c) for lung cancer data set.

6 Discussion

In this work, using a treed LFAM we infer subclonal genotypes structure for mutation pairs, their cellular proportions and the phylogenetic relationship among subclones. This is the first attempt to generate a subclonal phylogenetic structure using mutation pair data. We show that more accurate inference can be obtained using mutation pairs data compared to using only marginal counts for single SNVs. The model can be easily extended to incorporate more than two SNVs. Another way of extending the model is to encode mutation times inside the length of the edges of phylogenetic tree.

The major motivation for accurate estimation of heterogeneity in tumor is personalized medicine. The next step towards this goal is to use varying estimates of subclonal proportions across patients to drive adaptive treatment allocation.

Currently the heterogeneity is measured mostly with SNV and CNA data. However, structural variants (SVs) like deletion, duplication, inversion, translocation and other

large genome rearrangement arguably provide more accurate [Fan et al. (2014)] VAF estimation, which is the key input for characterizing the heterogeneity. Therefore incorporation of SVs into the current model could significantly improve the outcome of tumor heterogeneity analysis. Recently, in Brocks et al. (2014) the authors attempted to explain the intratumor heterogeneity in DNA methylation and copy-number pattern by a unified evolutionary process. So the current genome centric definition of tumor heterogeneity could be extended by incorporation of methylation, DNA mutation, and RNA expression data in an integromics model.

Finally in the era of big data it is important to factor computation into the research effort, and build efficient computational models that could handle millions of SNVs. Linear response variational Bayes [Giordano et al. (2015)] or MAD-Bayes [Broderick et al. (2013); Xu et al. (2015)] methods could be considered as alternative computational strategies to tackle the problem.

References

- Adams, R. P., Z. Ghahramani, and M. I. Jordan (2010). Tree-structured stick breaking for hierarchical data. In *Advances in neural information processing systems*, pp. 19–27.
- Aparicio, S. and C. Caldas (2013). The implications of clonal genome evolution for cancer medicine. *New England journal of medicine* 368(9), 842–851.
- Bafna, V., D. Gusfield, G. Lancia, and S. Yooseph (2003). Haplotyping as perfect phylogeny: a direct approach. *Journal of Computational Biology* 10(3-4), 323–340.
- Bignell, G. R., C. D. Greenman, H. Davies, A. P. Butler, S. Edkins, J. M. Andrews, G. Buck, L. Chen, D. Beare, C. Latimer, et al. (2010). Signatures of mutation and selection in the cancer genome. *Nature* 463(7283), 893–898.
- Bonavia, R., W. K. Cavenee, F. B. Furnari, et al. (2011). Heterogeneity maintenance in glioblastoma: a social network. *Cancer research* 71(12), 4055–4060.
- Bozic, I., T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K. W. Kinzler, B. Vogelstein, and M. A. Nowak (2010). Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences* 107(43), 18545–18550.
- Brocks, D., Y. Assenov, S. Minner, O. Bogatyrova, R. Simon, C. Koop, C. Oakes, M. Zucknick, D. B. Lipka, J. Weischenfeldt, et al. (2014). Intratumor DNA methylation

- heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell reports* 8(3), 798–806.
- Broderick, T., B. Kulis, and M. Jordan (2013). MAD-Bayes: MAP-based asymptotic derivations from Bayes. In *Proceedings of The 30th International Conference on Machine Learning*, pp. 226–234.
- Carter, S. L., K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* 30(5), 413–421.
- Chipman, H. A., E. I. George, and R. E. McCulloch (1998). Bayesian CART model search. *Journal of the American Statistical Association* 93(443), 935–948.
- Dagum, L. and R. Menon (1998). OpenMP: an industry standard API for shared-memory programming. *Computational Science & Engineering, IEEE* 5(1), 46–55.
- Denison, D. G. T., B. K. Mallick, and A. F. M. Smith (1998). A Bayesian CART algorithm. *Biometrika* 85(2), pp. 363–377.
- Deshwar, A. G., S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris (2015). PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology* 16(1), 35.
- Fan, X., W. Zhou, Z. Chong, L. Nakhleh, and K. Chen (2014). Towards accurate characterization of clonal heterogeneity based on structural variation. *BMC bioinformatics* 15(1), 1.
- Fischer, A., I. Vázquez-García, C. J. Illingworth, and V. Mustonen (2014). High-definition reconstruction of clonal composition in cancer. *Cell reports* 7(5), 1740–1752.
- Gerlinger, M., A. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C. Santos, M. Nohadani, A. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P. Futreal, and C. Swanton (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med.* 366(10), 883–892.
- Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. Interface Foundation of North America.

- Giordano, R. J., T. Broderick, and M. I. Jordan (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems*, pp. 1441–1449.
- Gusfield, D. (1991). Efficient algorithms for inferring evolutionary trees. *Networks* 21(1), 19–28.
- Jiao, W., S. Vembu, A. G. Deshwar, L. Stein, and Q. Morris (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics* 15(1), 35.
- Lee, J., P. Müller, K. Gulukota, Y. Ji, et al. (2015). A Bayesian feature allocation model for tumor heterogeneity. *The Annals of Applied Statistics* 9(2), 621–639.
- Li, H. and R. Durbin (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14), 1754–1760.
- Marass, F., F. Mouliere, K. Yuan, N. Rosenfeld, and F. Markowetz (2017). A phylogenetic latent feature model for clonal deconvolution. *The Annals of Applied Statistics* 10(4), 2377–2404.
- Marusyk, A., V. Almendro, and K. Polyak (2012). Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer* 12(5), 323–334.
- McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20(9), 1297–1303.
- Miller, C. A., B. S. White, N. D. Dees, M. Griffith, J. S. Welch, O. L. Griffith, R. Vij, M. H. Tomasson, T. A. Graubert, M. J. Walter, et al. (2014). SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS computational biology* 10(8), e1003665.
- Miller, K. T., T. Griffiths, and M. I. Jordan (2008). The phylogenetic Indian buffet process: a non-exchangeable nonparametric prior for latent features. In *Proc. UAI*.
- Navin, N., A. Krasnitz, L. Rodgers, K. Cook, J. Meth, J. Kendall, M. Riggs, Y. Eberling, J. Troge, V. Grubor, et al. (2010). Inferring tumor progression from genomic heterogeneity. *Genome research* 20(1), 68–80.

- Nik-Zainal, S., P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna, et al. (2012). The life history of 21 breast cancers. *Cell* 149(5), 994–1007.
- Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science* 194(4260), 23–28.
- Oesper, L., A. Mahmoody, and B. J. Raphael (2013). THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol* 14(7), R80.
- O’Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B* 57, 99–138.
- Papaemmanuil, E., M. Cazzola, J. Boulton, L. Malcovati, P. Vyas, D. Bowen, A. Pellagatti, J. Wainscoat, E. Hellstrom-Lindberg, C. Gambacorti-Passerini, et al. (2011). Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *New England Journal of Medicine* 365(15), 1384–1395.
- Pe’er, I., T. Pupko, R. Shamir, and R. Sharan (2004). Incomplete directed perfect phylogeny. *SIAM Journal on Computing* 33(3), 590–607.
- Raphael, B. J., J. R. Dobson, L. Oesper, and F. Vandin (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome medicine* 6(1), 1.
- Roth, A., J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, and S. P. Shah (2014). PyClone: statistical inference of clonal population structure in cancer. *Nature methods* 11(4), 396–398.
- Sengupta, S., K. Gulukota, Y. Zhu, C. Ober, K. Naughton, W. Wentworth-Sheilds, and Y. Ji (2016). Ultra-fast local-haplotype variant calling using paired-end DNA-sequencing data reveals somatic mosaicism in tumor and normal blood samples. *Nucleic acids research* 44(3), e25–e25.
- Sengupta, S., J. Wang, J. Lee, P. Müller, K. Gulukota, A. Banerjee, and Y. Ji (2015). BayClone: Bayesian nonparametric inference of tumor subclones using NGS data. In *Proceedings of The Pacific Symposium on Biocomputing (PSB)*, Volume 20, pp. 467–478.
- Stephens, P. J., P. S. Tarpey, H. Davies, P. Van Loo, C. Greenman, D. C. Wedge, S. Nik-Zainal, S. Martin, I. Varela, G. R. Bignell, et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature* 486(7403), 400–404.

- Strino, F., F. Parisi, M. Micsinai, and Y. Kluger (2013). TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Research* 41(17), e165.
- Van Loo, P., S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, et al. (2010). Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences* 107(39), 16910–16915.
- Varela, I., P. Tarpey, K. Raine, D. Huang, C. K. Ong, P. Stephens, H. Davies, D. Jones, M.-L. Lin, J. Teague, et al. (2011). Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* 469(7331), 539–542.
- Xu, Y., P. Müller, Y. Yuan, K. Gulukota, and Y. Ji (2015). MAD Bayes for tumor heterogeneity – feature allocation with exponential family sampling. *Journal of the American Statistical Association* 110(510), 503–514.
- Zare, H., J. Wang, A. Hu, K. Weber, J. Smith, D. Nickerson, C. Song, D. Witten, C. A. Blau, and W. S. Noble (2014). Inferring clonal composition from multiple sections of a breast cancer. *PLoS computational biology* 10(7), e1003703.
- Zhou, T., P. Müller, S. Sengupta, and Y. Ji (2017). PairClone: A Bayesian subclone caller based on mutation pairs. *arXiv preprint arXiv:1702.07465*.